CAPS-DB: a structural classification of helix-capping motifs

Segura, Joan; Oliva, Baldomero; Fernandez-Fuentes, Narcis

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gkr879

Publication date:
2011

Citation for published version (APA):
Segura, J., Oliva, B., & Fernandez-Fuentes, N. (2011). CAPS-DB: a structural classification of helix-capping motifs. *Nucleic Acids Research*, *40*(D1), D479-D485. https://doi.org/10.1093/nar/gkr879

Published online 22 October 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D479–D485***
doi:10.1093/nar/gkr879

CAPS-*DB*: a structural classification of helix-capping motifs

Joan Segura[1], Baldomero Oliva[2] and Narcis Fernandez-Fuentes[1,*]

[1]Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, University of Leeds, St. James's University Hospital, Leeds, LS9 7TF, UK and [2]Structural Bioinformatics Lab (GRIB), Universitat Pompeu Fabra-IMIM, Barcelona Research Park of Biomedicine (PRBB), 08003-Barcelona, Catalonia, Spain

Received August 3, 2011; Revised and Accepted September 29, 2011

ABSTRACT

The regions of the polypeptide chain immediately preceding or following an α-helix are known as Nt- and Ct cappings, respectively. Cappings play a central role stabilizing α-helices due to lack of intrahelical hydrogen bonds in the first and last turn. Sequence patterns of amino acid type preferences have been derived for cappings but the structural motifs associated to them are still unclassified. CAPS-*DB* is a database of clusters of structural patterns of different capping types. The clustering algorithm is based in the geometry and the (ϕ–ψ)-space conformation of these regions. CAPS-*DB* is a relational database that allows the user to search, browse, inspect and retrieve structural data associated to cappings. The contents of CAPS-*DB* might be of interest to a wide range of scientist covering different areas such as protein design and engineering, structural biology and bioinformatics. The database is accessible at: http://www.bioinsilico.org/CAPSDB.

INTRODUCTION

The first and last turn of α-helices lack intrahelical hydrogen-bonds requiring its Ct and N-ends to be stabilized by cappings to avoid the fray. Capping motifs were first described by Richardson and Rose in 1988 (1,2) as formed by interactions of the last residues of the helix with the closest residues of the polypeptide sequence. Subsequent works extended the classification of helical Ct- (3,4) and Nt cappings (5–9) and terms such as the 'Schellman', capping-box, big-box or αL capping motifs were introduced. A more systematic classification of cappings was presented by Aurora and Rose (10) by analysing a large data set of protein structures classifying helix-capping motifs according to specific patterns of hydrogen bonding and hydrophobic interactions found at or near the ends of helices in both proteins and peptides. Consequently, cappings play an important role in the stability (11–16) and function of proteins (17). Its functional relevance is in general associated with structure stabilization of a protein or fold (18), which occasionally may imply other type of associated disorders or diseases [e.g. in a prion-protein stability (19) or diabetes mellitus produced by a misfolded transcription factor (20)].

Although there has been extensive studies showing the sequence preferences and atomic interactions in cappings, there has not been a systematic classification of the structural patterns associated to cappings. We have used the definition of the topological features of the classification of loops to cluster the conformations of the polypeptide chain forming the capping in order to automatically extend the pre-genomic era classification. As a result a novel and automated structural classification and database of helix cappings, CAPS-*DB*, is presented. Cappings were extracted from high-quality protein structures and structurally clustered based on geometry and conformation (ϕ–ψ angles). The clustering process does not require the structural alignment of cappings but a two-step process were both geometry and conformation are compared. This procedure allows for a post-clustering quality check based in Root Mean Square Deviation (RMSD) values. The entire process, from extraction to clustering and archiving is fully automated, which results on CAPS-*DB* being updated on a regular basis. The current version of CAPS-*DB*, v.2.0, contains 16 195 Nt cappings and 16 188 Ct cappings extracted from 3848 protein structures that have been classified into 905 clusters. CAPS-*DB* features a clear and intuitive web interface that allows users to search, browse and retrieve data easily and conveniently.

*To whom correspondence should be addressed. Tel: +44 (0) 113 3438614; Fax: +44 (0) 113 3438601; Email: N.Fernandez-Fuentes@leeds.ac.uk
Present address:
Fernandez-Fuentes, Institute of Biological, Environmental and Rural Science, Aberystwyth University, Goggerddan Campus, Aberystwyth SY23 3EB, United Kingdom, Email: nff@aber.ac.uk.

DEFINITION OF HELIX-CAPPING MOTIFS AND GEOMETRY

The definition of capping used in this work fall in the short and medium-range capping interaction categories proposed by Aurora and Rose (10). Thus, the residues before a helix define an Nt capping where the first residue of the capping interacts with residue(s) in the first turn of the helix. Conversely, a Ct capping is defined by the portion of the polypeptide chain between the end of a helix and the last residue that interacts with any group in the last turn of the helix (Figure 1A).

The concept of geometry derives from our previous work in loop structure classification (21,22). In the case of cappings, the geometry characterizes the topology and is defined by two internal coordinates: a distance and an angle. The distance D is defined by the Euclidian distance between the first (Nt capping) or last (Ct capping) C-α of the α-helix and the first (Nt capping) or last (Ct capping) C-α sequentially adjacent that is within atomic interaction distance of any of the residues in the first (N1, N2, N3, N4) or last (C-1, C-2, C-3, C-4) turn of the α-helix (nomenclature for helix residues as defined by Aurora and Rose (10)). The delta angle is defined by the angle between the vector defined by the moment of inertia M of the α-helix and the distance vector D (Figure 1A). The range of parameter D would depend on the number of residues that form the capping being typically <16 Å while the delta angle ranges between 0 and 180°.

CLUSTERING ALGORITHM

The clustering algorithm operates at two different levels: geometry and conformation. On the first stage, cappings are merged into groups that share the same geometry. Two cappings would have the same geometry if Δ(D, delta) values belongs to the two dimensional semi-open interval $I = [(0,0),(4,45)]$. The partition of the geometrical space was optimized via a comprehensive exploration of different bin definitions. During the second stage, cappings with the same geometry are clustered based on a conformation similarity score. The conformation of the residues within the capping is encoded by assigning a conformation code that depends on the ϕ–ψ angles according to the partition of the Ramachandran space showed in Figure 1B. Thus, the conformation of a given capping is encoded by a string of characters each of which described a precise region of the (ϕ–ψ)-space. The clustering method is based on a density search algorithm (23). Because no structural superposition of cappings is required during the clustering, then a structural similarity measure, the RMSD of main chain atoms, is used to assess the quality of the clusters (see next).

QUALITY OF CLUSTERS

The clustering algorithm is based on geometry and density search on the (ϕ–ψ)-space, is therefore independent of structural alignment between cappings. The RMSD within and between cappings of the same and different

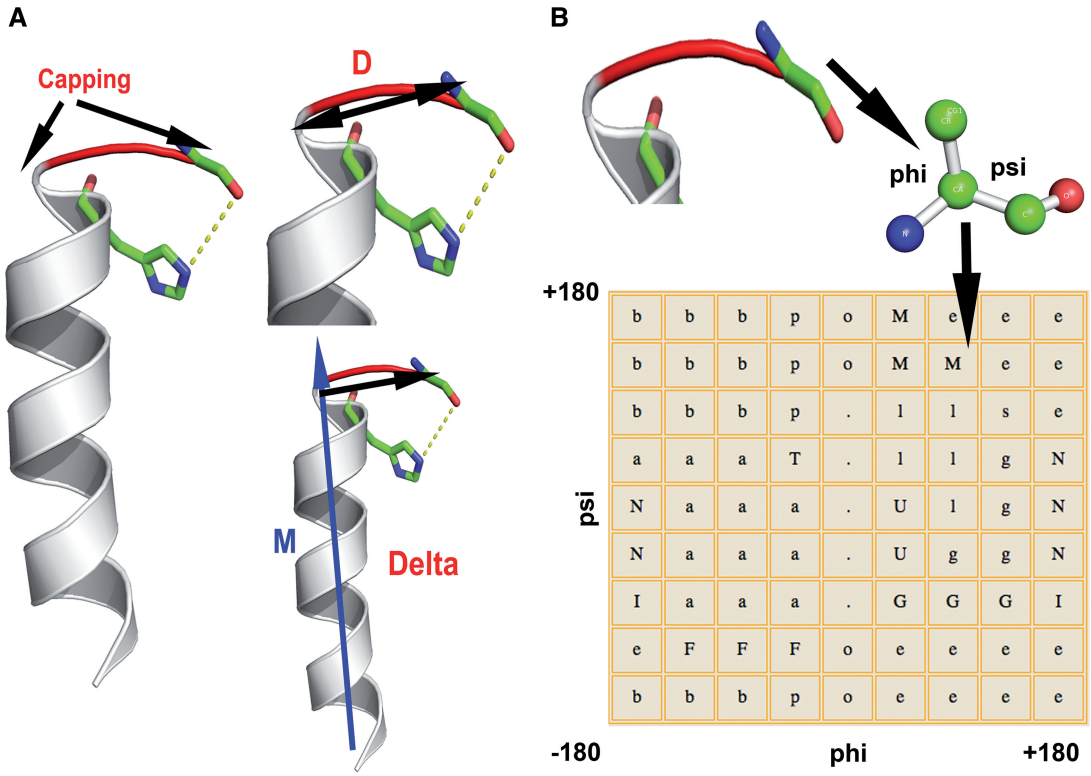

Figure 1. Definition of capping, geometry of cappings and partition and encoding of phi/psi angles. (**A**) Example of a Nt capping and geometry descriptors. (**B**) Encoding of the capping conformation as a string of characters using the partition of phi/psi space. The most accessible regions of the phi/psi space encoded as follow: a: α-helix; l: left-handed α-helix; b: β-strand; p: β-prolin; g: γ-helix; e: ε.

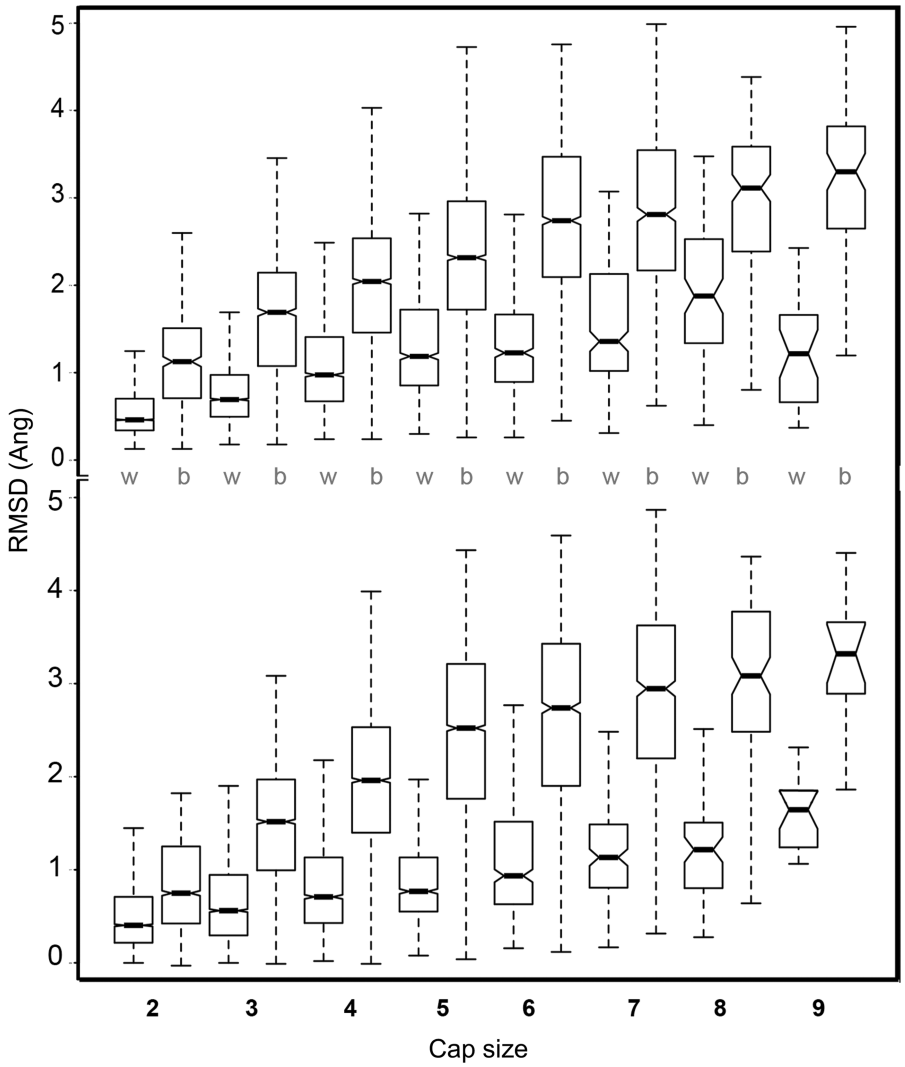

Figure 2. Box plots of RMSD values as a function of capping size within (w) and between (b) clusters. The central horizontal line in the box marks the median and the box edges the first and third quartile; errors bars show minimum and maximum values. Box plots for Nt- and Ct cappings are shown in the upper and lower plot, respectively.

clusters is then used to assess the quality of the clustering. Figure 2 shows the distribution of RMSD (main chain) upon structural alignment of cappings that belong to the same (w) or different clusters (b). Amongst cappings that have the same size and for all capping sizes the distribution of RSMD values is significantly smaller for those cappings that belong to the same cluster (w) to those that belong to different cluster (b). This result proves that the clustering algorithm is indeed grouping structurally related cappings. The clustering relies on comparison of strings values (i.e. geometry and conformation) and do not require complicate and computing intensive operations such calculation of rotation/translation matrices (e.g. performing structure superposition), and thus is very fast and well suited for large-scales analysis. Being the clustering of capping the central element, then keeping CAPS-*DB* up-to-date will not require neither extensive nor intensive computing, and thus ensuring its long-term viability.

DATABASE IMPLEMENTATION AND CONTENTS

CAPS-*DB* comprises two major components: a relational database for data storage and management and a web application to interface the database. Data is stored in a relational MySQL database whose design was optimized to provide a fast and optimal access to the information. It makes extensive use of master and internal keys and cross-references between tables. The MySQL server runs in a dedicated computer that also mirror all external databases that are required during the updating and annotation process, e.g. PDB (24).The web application runs on an Apache web-server hosted on a Red Hat® enterprise Linux operating system. CGI Perl, JavaScript and DBI-DBD modules are used to interface and access the database. Web pages resulting from queries and sequence searches are generated *on the fly*, i.e. are dynamic, thus ensuring up-to-date information is available to users. The web-site includes a Jmol applet (http://jmol.sourceforge.net) to visualize protein structures and a BLAST (25)

search engine to perform sequences searches (see *Retrieving information* section). Finally, documentation and explanations about contents, clustering and the use of CAPS-*DB* is provided to users in the help section of the website.

The current version of CAPS-*DB*, v.2.0, classifies over 30 000 cappings extracted from 3848 protein structures and grouped in 905 clusters. The process of extraction and classification of cappings was performed as follow. An initial set of non-redundant protein structures was derived from the protein databank (PDB) (24) using PISCES (26). The parameters used to generate the initial set were: crystal resolution better that 2.0 Å, a maximum *R*-factor of 0.25, minimum chain length of 40 residues and a sequence cut-off of 25%. This initial set was subjected to further quality filters that included the checking for non-standard amino acids that were discarded with the exception of Se-Met that were converted into Met, atoms with alternative locations (only the first rotamer was kept) and amino acids with insertion codes that were renumbered if non-superposable. Missing main-chain and side-chain atoms were added using Maxsprout (27) and Scwrl 4.0 (28), respectively. The secondary structure of proteins was assigned using DSSP (29) and the atomic interaction of capping-helices were defined using CSU (30). Finally, cappings and flanking helices were extracted and clustered using the approach described before.

QUERYING AND RETRIEVING DATA

There are different basic approaches to query and retrieve data from CAPS-*DB*. The first approach is by simply browsing the ontology of the classification (Figure 3). The first level is the cappings type: Nt- or Ct cappings. The second level is the capping size, i.e. the number of residues in the capping. Below capping size there are two more levels that relate to the geometry of the capping: D and δ angle. Clusters are the lower level and present the cappings that are structurally equivalent and other related information (see next *Cluster information*). The second approach to query CAPS-*DB* is by simply providing the PDB identification code of the protein of interest in the text box embedded in the top menu (Figure 3). If the protein is classified in CAPS-*DB*, the server will return a list of clusters and the associated cappings.

Users can also query CAPS-*DB* by doing a sequence search using a BLAST (25) engine implemented in the server (Figure 3). A sequence identity cut-off value and substitution matrix can be selected in the advanced option menu. The server will return a list of target proteins sorted by the BLAST scores and alignments can be inspected by following the relevant links (Figure 3). Finally, CAPS-*DB* also features an advanced search engine that allows more complex and elaborated queries that includes: searching for capping type, capping size or capping geometry; searches using free text or keywords (i.e. reductases), Unitprot accession number (13), or PubMed identifier; and any combination of the aforementioned queries.

CLUSTER INFORMATION

The clusters of the cappings are presented in individual web pages that consist of two main components: a general information section and a table with alignment of the cappings (Figure 4).

The general section provides the following information: the type of capping (Nt or Ct), cappings size and the number of capping-motifs included in the cluster; the consensus geometry; the sequence, Ramachandran information on (ϕ,ψ)-conformational space, and secondary structure consensus; the average sequence identity among cappings included in the cluster; the average RMSD between capping regions (calculated using main-chain atoms); a PROSITE-like sequence pattern and the pattern entropy. The consensus sequences was derived from the aligned cappings selecting the amino acid type if conserved >90%, p for polar residues [D,E,H,K,N,Q,R,S,T,Y] or h for non-polar residues [A,C,F,G,I,L,M,PV,W,Y] if conserved in at least 75% of the sequences; 'X' otherwise. Ramachandran and secondary structure consensus were derived similarly. The PROSITE-like sequence pattern was derived from the weighted [as in Henikoff and Henikoff (31)] amino acid frequencies derived from the alignment where amino acids type are shown in lower case if having an individual frequency between 75% and 90% and upper case if >90%; 'x' otherwise. The pattern entropy was calculated using the AL2CO program (32) with default parameters.

Another important element in the general information section is the so-called *contact patterns matrix*. This is a graphical matrix that shows the proportion of cappings in the cluster that have a given interaction pair, e.g. capping-motifs with an atomic interactions between C′ and C4 residues [nomenclature of residues adopted from Aurora and Rose (10)]. This matrix provides a visual representation of the atomic interactions that are more important and conserved in the cappings, and thus interactions patterns that are more prevalent among capping-motifs. Detailed information about the atomic interaction types including structural visualization is also provided (see 'Conclusion' section).

Cappings-motifs belonging to a cluster are presented in a table that includes information about protein structures of origin (PDB code, chain, start residue and end residues; numbering as in the PDB file), sequence, secondary structure [as defined by DSSP (29)] and Ramachandran. The contents of the table and other additional files can be downloaded by following the relevant links. These include the superposed coordinates in compressed format (.gz); a PSSM profile; and a tab-delimited plain text file detailing the atomic interactions between cappings and helix termini. Finally, a Jmol applet allows the visualization of the structure of cappings by selecting them from a clickable list (Figure 4), including the atomic interactions between cappings and helix.

CONCLUSION

Here we present an automatic structural classification of cappings. The classified cappings presented in this work

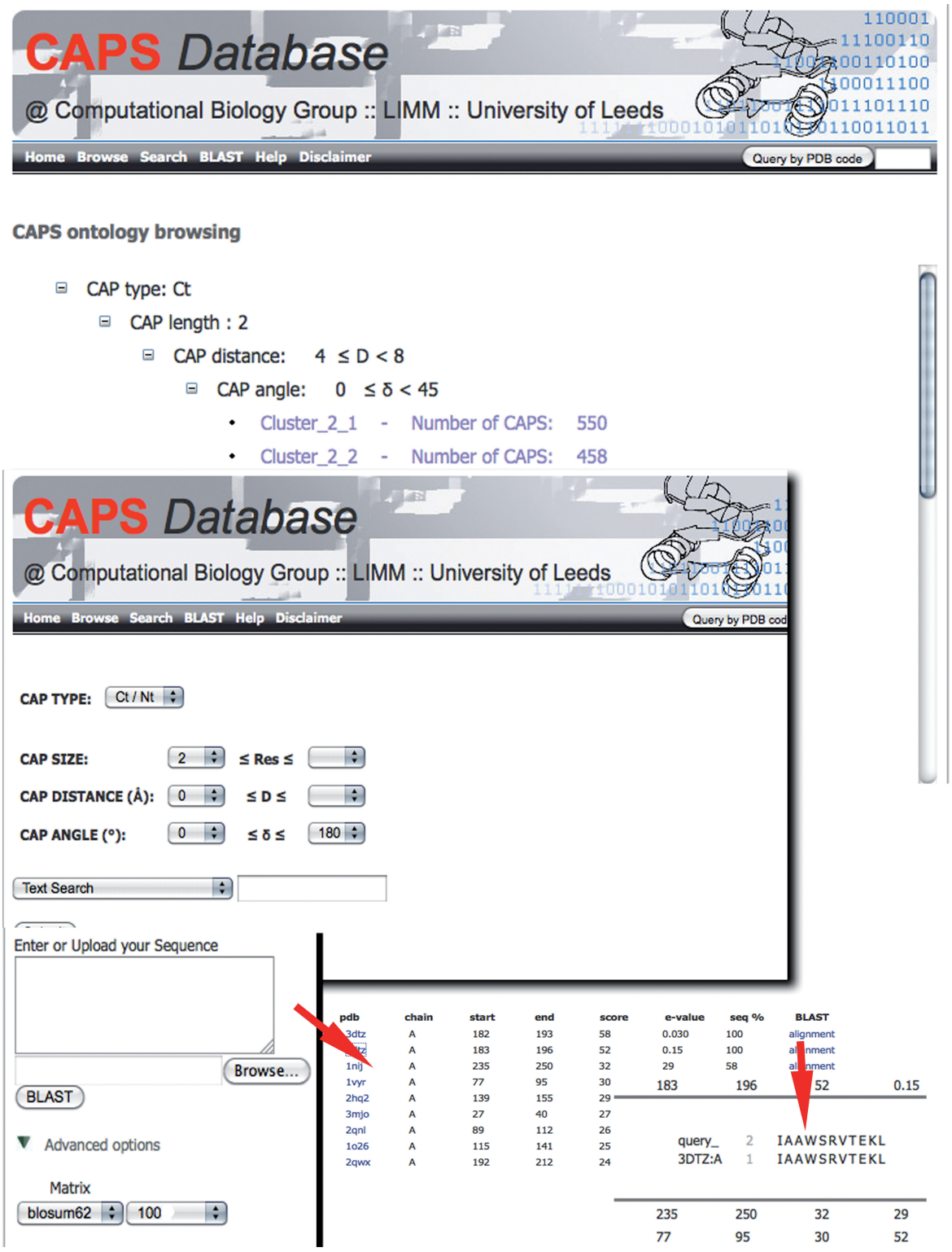

Figure 3. Screenshot of a CAPS-*DB* ontology browser, search web page and BLAST sequence search engine including a BLAST output and alignment.

have been stored in a MySQL relational database named CAPS-*DB*. The clustering approach can recapitulate structural motifs for classical cappings such as the big-box, αL-motif or Schellman (10). Users can browse among 905 types of different capping conformations extracted from a non-redundant set of 3848 protein structures. Additionally, the database can be queried using a BLAST sequence search, functional-keyword search

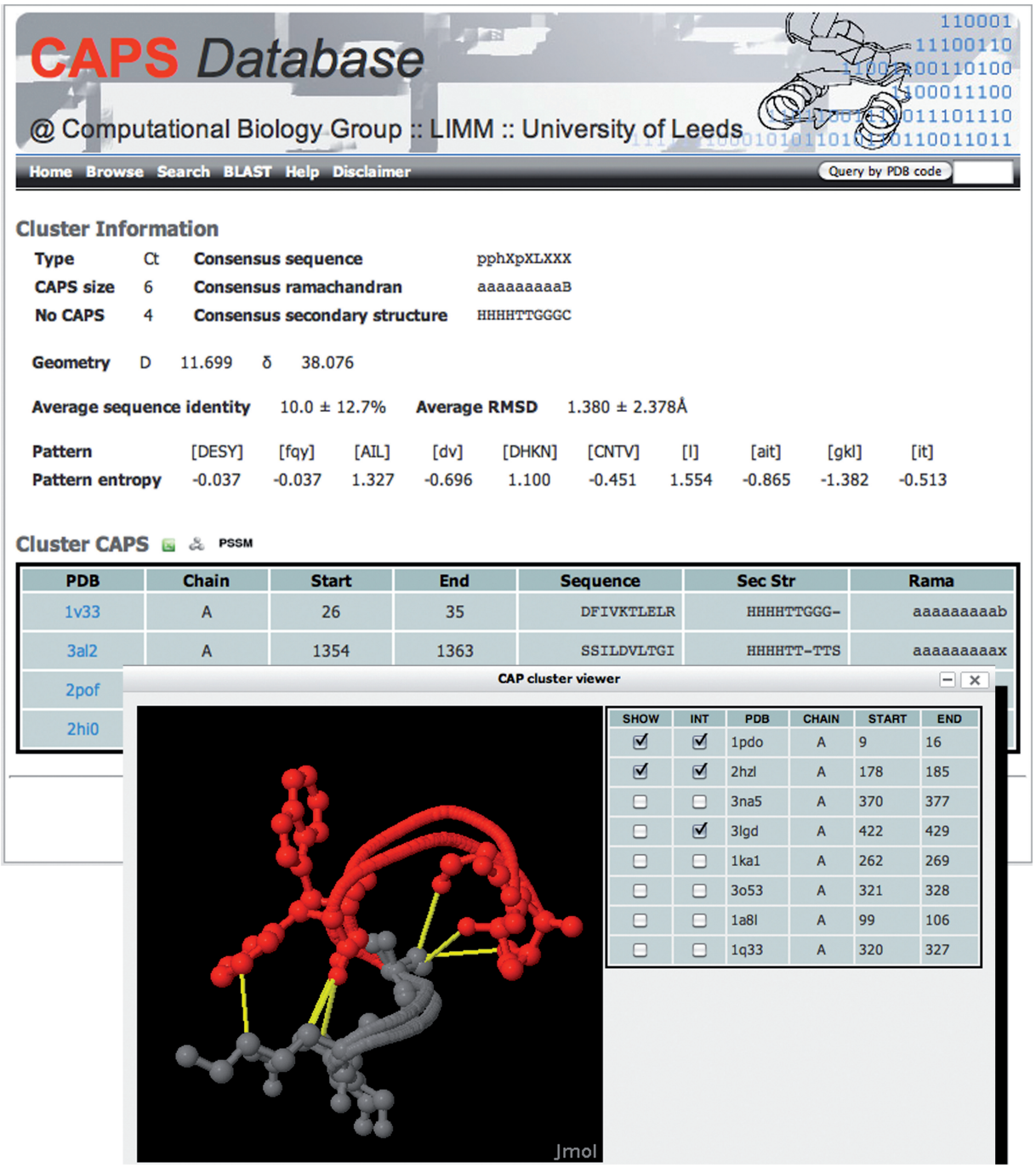

Figure 4. Screenshot of CAPS-*DB* cluster information web page and Jmol applet.

or any of the topological features used for the clustering. The information retrieved from the database accounts for the sequence profile and relevant conservation of specific residues. Finally, links are provided to the original PDB structures and visualization of cappings using Jmol.

There are a number of areas where the data compiled in CAPS-*DB* can be applied. In comparative modelling and structural biology, the information classified in CAPS-DB can be used to define the boundaries of helices by comparing the sequence of helix to these classified in CAPS-*DB*. It can be also an important resource in protein design to improve the stability of helices by optimizing Ct and Nt cappings. Finally, sequence profiles derived from structural clusters can be used to improve secondary structure prediction. Summarizing, CAPS-*DB* is a useful tool to predict structural and functional local-features and apply them to optimize in models, refine in crystals, design, construct and understand helix boundaries in proteins.

ACKNOWLEDGEMENTS

N.F.F. thanks Dr Gendra for critical reading and insightful comments to the manuscript, and Ms Martina and Ms Daniela G Fernandez for continuing inspiration and motivation.

FUNDING

Research Councils United Kingdom Academic Fellow scheme (to N.F.F.); internal scholarship awarded by the Leeds Institute of Molecular Medicine (to J.S.M.); MICINN and FEDER (BIO2011-22568) to B.O. Funding for open access charge: Research Councils United Kingdom (RCUK)

Conflict of interest statement. None declared.

REFERENCES

1. Presta,L.G. and Rose,G.D. (1988) Helix signals in proteins. *Science*, **240**, 1632–1641.
2. Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**, 1648–1652.
3. Milner-White,E., Ross,B.M., Ismail,R., Belhadj-Mostefa,K. and Poet,R. (1988) One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. *J. Mol. Biol.*, **204**, 777–782.
4. Aurora,R., Srinivasan,R. and Rose,G.D. (1994) Rules for alpha-helix termination by glycine. *Science*, **264**, 1126–1130.
5. Stickle,D.F., Presta,L.G., Dill,K.A. and Rose,G.D. (1992) Hydrogen bonding in globular proteins. *J. Mol. Biol.*, **226**, 1143–1159.
6. Seale,J.W., Srinivasan,R. and Rose,G.D. (1994) Sequence determinants of the capping box, a stabilizing motif at the N-termini of alpha-helices. *Protein Sci.*, **3**, 1741–1745.
7. Munoz,V., Blanco,F.J. and Serrano,L. (1995) The hydrophobic-staple motif and a role for loop-residues in alpha-helix stability and protein folding. *Nat. Struct. Biol.*, **2**, 380–385.
8. Dasgupta,S. and Bell,J.A. (1993) Design of helix ends. Amino acid preferences, hydrogen bonding and electrostatic interactions. *Int. J. Pept. Protein Res.*, **41**, 499–511.
9. Harper,E.T. and Rose,G.D. (1993) Helix stop signals in proteins and peptides: the capping box. *Biochemistry*, **32**, 7605–7609.
10. Aurora,R. and Rose,G.D. (1998) Helix capping. *Protein Sci.*, **7**, 21–38.
11. Tripet,B. and Hodges,R.S. (2002) Helix capping interactions stabilize the N-terminus of the kinesin neck coiled-coil. *J. Struct. Biol.*, **137**, 220–235.
12. Bang,D., Gribenko,A.V., Tereshko,V., Kossiakoff,A.A., Kent,S.B. and Makhatadze,G.I. (2006) Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis. *Nat. Chem. Biol.*, **2**, 139–143.
13. Ermolenko,D.N., Thomas,S.T., Aurora,R., Gronenborn,A.M. and Makhatadze,G.I. (2002) Hydrophobic interactions at the Ccap position of the C-capping motif of alpha-helices. *J. Mol. Biol.*, **322**, 123–135.
14. Makhatadze,G.I. (2005) Thermodynamics of alpha-helix formation. *Adv. Protein Chem.*, **72**, 199–226.
15. Thomas,S.T., Loladze,V.V. and Makhatadze,G.I. (2001) Hydration of the peptide backbone largely defines the thermodynamic propensity scale of residues at the C' position of the C-capping box of alpha-helices. *Proc. Natl Acad. Sci. USA*, **98**, 10670–10675.
16. Thomas,S.T. and Makhatadze,G.I. (2000) Contribution of the 30/36 hydrophobic contact at the C-terminus of the alpha-helix to the stability of the ubiquitin molecule. *Biochemistry*, **39**, 10275–10283.
17. Rezvanpour,A., Phillips,J.M. and Shaw,G.S. (2009) Design of high-affinity S100-target hybrid proteins. *Protein Sci.*, **18**, 2528–2536.
18. Koscielska-Kasprzak,K., Cierpicki,T. and Otlewski,J. (2003) Importance of alpha-helix N-capping motif in stabilization of betabetaalpha fold. *Protein Sci.*, **12**, 1283–1289.
19. Iovino,M., Falconi,M., Petruzzelli,R. and Desideri,A. (2001) Role of the helix capping in the stability of the mouse prion (180-213) segment: investigation through molecular dynamics simulations. *J. Biomol. Struct. Dyn.*, **19**, 237–246.
20. Narayana,N., Phillips,N.B., Hua,Q.X., Jia,W. and Weiss,M.A. (2006) Diabetes mellitus due to misfolding of a beta-cell transcription factor: stereospecific frustration of a Schellman motif in HNF-1alpha. *J. Mol. Biol.*, **362**, 414–429.
21. Oliva,B., Bates,P.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814.
22. Fernandez-Fuentes,N., Oliva,B. and Fiser,A. (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res.*, **34**, 2085–2097.
23. Everitt,B. (1974) *Cluster Analysis, Chapt. 3*. Heineman Educational Books Ltd, London.
24. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 899–907.
25. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403.
26. Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
27. Holm,L. and Sander,C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.*, **218**, 183.
28. Krivov,G.G., Shapovalov,M.V. and Dunbrack,R.L. Jr (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.
29. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
30. Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
31. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weigths. *J. Mol. Biol.*, **243**, 574–578.
32. Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.